EMOTION RECOGNITION IN STANDARD SPOKEN ASSAMESE LANGUAGE USING SUPPORT VECTOR MACHINE AND ENSEMBLE MODEL

Nupur Choudhury

Research Scholar, Department of Computer Science & Engineering, Assam Don Bosco University, Azara, Guwahati, Assam, 781017, India
nupur.choudhury@dbuniversity.ac.in

Uzzal Sharma

Assistant Professor, Department of Computer Applications, Assam Don Bosco University, Azara, Guwahati, Assam, 781017, India
uzzal.sharma@dbuniversity.ac.in

Abstract

This paper deals with the emotion recognition of standard spoken Assamese language by using Support Vector Machine and Ensemble Model. An Ensemble model has been generated using random forest classifier and rotation forest classifier. These efficient classifiers were combined and compared against the standard Support Vector Machine and were found that the accuracy of Ensemble algorithms when 4 or more features are clubbed together is more than the standard Support Vector Machine results. A total of 7 emotions were recognized which included calm, neutral, anger, sad, happy, surprise and disgust. All these 7 emotions were tested for a variety of combination and different results were generated for each of them. This paper reports the findings where the Rotation Forest Classifier shows greater accuracy which is around 94.40% when combination of features are used than SVM or Random Forest Classifiers.

***Keywords*: Support Vector Machine, Machine learning, emotion recognition, Random Forest, Rotation Forest, classification etc.**

1. Introduction

Emotion detection in speech focuses on the identification of the physical state or the emotional state of the mind of human being utilizing their own voices or speech signals. These are also known as paralinguistic aspect of the speech signal where the speaker's emotional aspects are included in the speech. This emotional state necessarily does not alter the linguistic aspect but provides information or feedback relating to many aspects. A primary approach for this research is required due to increase in the number of suicide cases which is growing at an alarming rate in the entire world. Scientists have been into extensive research into finding the clue behind this aspect which have a prominent pointer towards a common emotion that has been the final cause of this major incidents-a derived emotion known as depression. According to the researchers and psychologists, the main reason behind this intentional accident although unknown but still can be estimated from heuristic approaches that reveals failure in examinations, relationships poverty etc. which might have been a reason for the sudden decision of such grievance and leads into taking the ultimate step. As such doctors and scientists are fighting to find a solution for detecting the emotions that might be going on in individual's thought process and look out for loopholes and depression if any so that this incident can be brought down to a certain level before completely eradicating it. While depression being just one aspect a lot of other primary and derived emotions also plays a major role in determining the mental stability of a person and various psychological deficiencies could be addressed in due time. Over 6,500 different languages are spoken all over the world out of which India alone has over 150 different localized languages. Several researches in various spoken languages like Caucasian, Sturim [27], German, Schuler [29], Berlin, Spanish, Swedish, German, English, Dutch, Chinese, Ververidis [28] etc. to name a few of the languages that has been utilized as a database to realize the effort. Emotion recognition for localized spoken languages is still posing a challenge when it comes to bringing down the number of depression related incidents in the region. Hence there is a need to address this problem from the social and psychological health of the youngsters and the individuals who might fall as a victim to this and can be saved from further suicidal activities. This paper gives a brief idea about the approach that would be serving as an effort to detect the various emotions

viz. calm, neutral, anger, sad, happy, surprise and disgust occurring in individuals speaking standard Assamese language.

A wide range of solutions involving Natural Language Processing solutions like Chatbot, Speech based systems needs speech input to work. Generally, the traditional procedure is to initially convert the speech signals to relevant text using the Automatic Speech Recognition (ASR) after which the conventional learning methodologies are applied for classification of the emotions. Kim et. al [2] in their research made use of CNN on previously trained word vectors for training and classification on sentence level and received good results in several benchmarks. Zhang et. al. made use of CNN character level parameters for classification related to converted text and received similar results as compared to conventional classification models that includes n-grams with TF-IDF variants, bag-of-words, Recurrent Neural Network (RNN) and ConvNets. In the present scenario the improved interaction and personalization of computers has made Human Computer Interaction better as they are increasing their efficiency in the prediction of emotional state of a human or speaker which leads to identifying various meaning related to a particular word for different context. ASR uses probabilistic language models and acoustics [3] and their speaker related variations where the speech features are not dependent on the speakers which could work for various applications but often proves less useful in context of the emotions intended to have the correct functionality. Efficient ASR generates highly accurate results but generally loses a huge amount of information related to emotions from the speech samples. This drawback has opened up a new area for research which is Speech-based Emotion Recognition (SER) during the last few years. Humans naturally express their emotions using speech due to which most of the modern-day applications work with emotion recognition objective work with speech signals. The traditional approach for SER is to extract various features from speech signals like the spectral features, frequency, pitch, formants and features related to energy and then apply classification in order to predict various emotions. [4][5]. Traditional classifications involve Hidden Markov Model (HMM)[8],Gaussian Mixture Model(GMM)[10], Bayesian Network Model[6][7], Support Vector Machine(SVM), Ensembled Models etc. There has been significant contributions from the field of Deep Learning in Natural Language Understanding since the last 10 years. For the last 10 years Natural Language Understanding has been deeply impacted using Deep Learning Techniques. Kim et al.[12] and Zheng et. al[13] proposed Deep Belief Networks(DBN) which demonstrated comparatively better performance over baseline models [2] [3] which do not use Deep learning and proves that higher order nonlinear associations are a better option for option recognition. Han et al. [14] worked on a Deep Neural Network-Extreme Learning Machine (ELM) that makes use of features from utterance by using segment level probability distribution in addition to one hidden layer neural network to detect emotions related to utterances with minimum accuracy. Fayek et al. [15] utilized deep hierarchy-based architectures, augmented data and a DNN regularization regarding SER. On the other hand, Zheng et al. [9] made use of spectrograms using Deep CNN. Vladimir et al. [10] worked with DNN on a series of acoustic features that are calculated using smaller speech intervals combined with a CTC loss function that is based on probability that took into consideration longer utterances which contains emotional as well as non-emotional segments with improved accuracy-based recognition. Lee et. al[5] made use of a bidirectional LSTM based model on the IEMOCAP[18] dataset for training the features and received an accuracy for emotion recognition of 62.8% that proved better to a great extent as compared to DNN-ELM. Recently the researchers in this domain are analyzing the utilization of multimodal features for emotion recognition. Tzirakis et al. [20] experimented with an SER system which makes use of auditory as well as visual modalities in order to capture emotion based contents from a wide variety of speaking styles. . Zadeh et al. [21] worked with a Tensor Fusion Network which has the capability of learning intra and inter modality completely and is good for online version of volatile languages. Ranganathan et al. [22] Convolutional Deep Belief Networks (CDBN) which uses salient expression features that are multimodal in nature in order to achieve better accuracies. Random forest proposed by L. Brieman [2] during 2001 is based on the begging principle. RF has been generating a new paradigm in every sector of machine learning and its applications including emotion recognition and Signal Processing applications. It has been found as a very powerful classification algorithm and it is affected by overfitting which is found extremely effective in handling higher dimensional data it has a better classification accuracy and model developing time in comparison to bagging and boosting. Its performance is not affected by size of the training data and the noise involved but is the sampling design create a deeper impact. RF is being applied successfully in various sectors of textual mining, regression, health sector etc. The significant part of RF is to selecting the base classifiers obvious data sets that are used as benchmarks. Rotation forest[30] is a new addition in existing Ensembled classification algorithms and it is observed to perform in a comparable manner with random forest on various data. It is considered as a better algorithm then random forest, boosting and bagging. However, it is not computationally economical and unstable then read numbers.

In this paper we are experimenting on some powerful classifiers namely Support Vector Machine (SVM) and Ensembled Model which is a collection of advanced machine learning algorithms like logistic regression, random forest classifier and rotation forest classifier, on a completely new and self-generated dataset in Assamese Language for emotion recognition. The main contribution of the current work are as follows:

- Development of a dataset for Emotion Recognition in Assamese Language

- Experiment on the dataset for emotion classification using SVM and Ensembled model with Speech Features (MFCC, Fundamental Frequency(F0), Zero Crossing Rate(ZCR), Linear Predictive Coding(LPC), Log Energy and Pitch)

2. Methodology

The primary approach of this research is to analyse various machine learning algorithms and try and classify the 7 categories of emotions that a person might be able to express through vocal activities in standard spoken Assamese language and analyse the accuracy for each one of them as well as finding an effective classifier so that the accuracy could be maximized. Currently extensive work is being carried out for detecting significant emotions using the combination of classical algorithms for each of the emotions mentioned above. The most important challenge being the spoken language itself, one cannot predict if the same algorithms would provide similar result for multiple languages. Hence development and analysis for a very dataset itself becomes a very challenging as well as non-predictable process as the same techniques used by different other spoken languages might not yield similar results or any results at all

2.1. *Dataset Used*

The Dataset that was used for experiment was a self-generated dataset which involved male and female actors of Assamese origin. The dataset generated is simulated and is generated by 10 male and 15 female actors over a period of 3 months' time in a closed environment. The following are the details of the dataset: The Dataset and the validation experiment was done using human volunteers. The participants were provided with a consent letter with the necessary information required which were signed by them prior to any recording or experiments that might be performed with their assistance. Well informed consent in written form was obtained before performing any type of Experiment or recording individually from the participants. The participants gave their consent in written form for any publication related factors as case details. Every individual and data from them were treated as a part of the experiment based on the Declaration of Helsinki. The methods of recording and related validation were done under supervision and primarily constituted of simulated speeches.

Development of the Dataset stimuli:
The dataset development is done in reference to Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[31].Twenty-five actors having near professional experience, studying and working in Assam Don Bosco University, Guwahati, Assam, India were involved in creating the stimulus (M = 26.0 years; SD = 3.75; age range = 20–33; 10 males and 15 females). Actors identified themselves as inhabitants of lower Assam, Upper Assam and plain areas. In order to be a part of the experiment the actors needed to have sound knowledge regarding the conventional spoken Assamese dialect derived from the written form and should have Assamese as their first language. They should be able to filter out the regional dialect and intonation while speaking the text provided to them. Five statements for 7 induced emotions per category was used which includes sentences spoken in Assamese. These induced emotions are normally referred as simulated, enacted, instructed or portrayed etc. The statements are around 10 syllables (maximum) in length. Some sample speeches include:

Emotion	Sentences
Neutral/calm	মই ভালে আছো । তোমাৰ কি খবৰ । অলপ ফুৰিব যাব উলাইছো ।
Happy	বাহ চাকৰী পাই গেলা। তোমাৰ বিয়াৰ কথা শুনি খুব ভাল লাগিছে । এইটো বিৰাট ভাল খবৰ ।
Sad	মোক যে বিৰাট দুখ দিলা । ই পৰীক্ষাত ফেইল কৰিলো । মনটো কিবা বেয়া লাগি আছে ।
Fear	ইমান ৰাতিকে কেনেকৈ যাও, ভয় লাগি আছে । আন্ধাৰ ৰুমটোৰ ভিতৰত কি কি আছে । মোৰ বেমাৰ হব যেন লাগি আছে ।

Angry	মোক হি গালি পাৰিছে । কি বনাইচ্ এইবোৰ তয়ে খাঁ । সি কিয় চিঞৰিব মোৰ উপৰত ।
Surprise	সেইটো চুন ধুনীয়া পখিলা । কি কোৱা হে, মই নাজানো । তুমি কেতিয়া আহিলা ।
Disgust	ইছ ইমান লেতেৰা মানুহজন । চেহ, তেণ্ডুলকাৰ আকৌ শূণ্যত আউট হল। ধেই নোৱাৰি খাব এইবোৰ ।

Table 1. The planning and control components.

The calm and neutral emotions played the condition baseline while the six other emotions constituted of fundamental emotion categories that are considered to be culturally universal. The dataset has around 1750 audio samples which were generated basically using 2 tones of the same statement by the same speaker. These 2 tones were selected keeping into consideration as higher and lower emotional expression for the same statement. Around 10 human validators were asked to validate the stimuli on a random basis. Necessary qualification to validate the speech samples and the stimuli essentially required Assamese as their first language. The validators made use of a choice based on a forced choice-respond framework where no escape option was provided. For testing purpose Unbiased hit rates (UHR) were calculated which had a proportion score of 0-1and generally give a smaller value in comparison to their correct values. However this UHR yield such result whenever there are accurate unbiased scores. UHR can be represented as:

$$UHR = Uncorrected\ hit\ rate * Differential\ accurcy \quad (1)$$

However, for the dataset validation the following was used

$$UHR_i = \frac{\sum_i (R_{Intended} = R_{Chosen})}{\sum_i R_{All}} \text{ x } \frac{\sum_i (R_{Intended} = R_{Chosen})}{\sum_i R_{All}}$$

$$UHR_i = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k} \quad (2)$$

Here I = ith stimulus and n is the number of samples which is responsible for each category and N is the total number of samples that were considered at a time for a speech signal. On generating the confusion matrix, the following were the results:

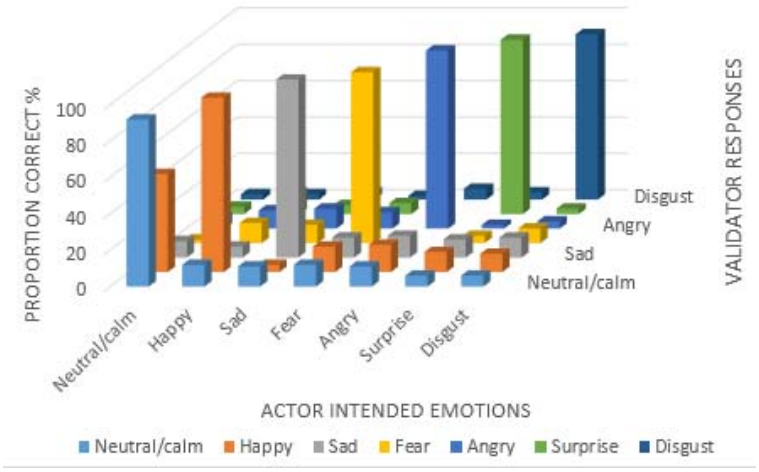


Fig. 1. Confusion Matrices of Emotional Validity.

Emotion Recognition System using Speech

The standard framework utilized for the emotion recognition system using speech that is used for experimentation is shown n Figure 2. Pre-processing of the signal is done by pre-emphasizing followed by framing and finally by windowing. In this work short term features like MFCC, Fundamental Frequency (F0), Zero Crossing Rate (ZCR), Linear Predictive Coding (LPC), Log Energy and Pitch are analyzed to generate effective results. Normalization of the features were done where they were calculated for every window of certain number of frames as well as fusion of the features were also carried out. Support Vector Machines (SVM) and Ensembled model were used as classifiers and the accuracy was analyzed.

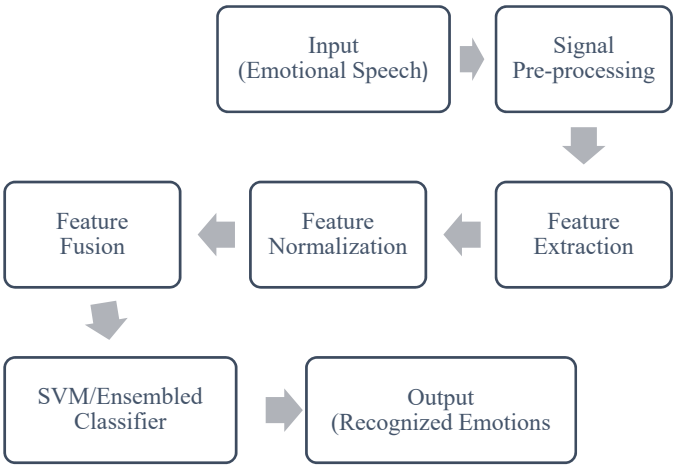


Fig. 2. Conventional Speech Processing steps.

A. Preprocessing of the signals

The primary operations which are used in pre-processing of the signals are pre-emphasis, framing of the signals and windowing of the signals

i. Pre-emphasis: In this work a Finite Impulse Response filter (FIR). It is a filter for flattening the spectrum of speech. The FIR used is given by equation 3.

$$H_{pre}(z) = 1 + a_{pre}z^{-1} \quad (3)$$

Where a_{pre} coefficient is set to 0.9375 for its efficient implementation in fixed point hardware.

ii. Framing: The audio signal is broken down into a series of frames where independent analysis of each frame is carried out and is depicted using one feature vector. Signal frames of length 25 msec is extracted from the signal that has already been filtered at every half or 1/3rd of the frame length.

iii. Windowing: A tapered window as been applied to each of the frames which is responsible for reducing the discontinuities of the signal at frame edges. A common windowing method known as Hamming window is applied as shown in (4).

$$W = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad ,0 \leq n \leq N-1 \quad (4)$$

B. Feature Extraction

Speech signals constitute of a variety of parameters which would reflect the emotion based characteristics. These wide variety of parameters are responsible for various emotional changes. Hence extraction of the features which expresses the emotions of a speech signal is a very important step. This paper deals with the characteristics of five short time features i.e. MFCC, Fundamental Frequency (F0), Zero Crossing Rate (ZCR), Linear Predictive Coding (LPC) and Log Energy. These are basically the acoustic features for emotion recognition using speech.

i. Fundamental frequency (F0): t is not processed on a linear scale rather it is processed often on a logarithmic scale since it matches the resolution of the auditory system of the human. The frequency range of 50-500Hz for speech signal (general human frequency for pitch) is used to generate the autocorrelation function [22].

ii. Energy: this is computed by considering the speech samples s_n in a particular time window by (5)

$$E_v = \sum_{n=1}^{N} {s_n}^2 \tag{5}$$

iii. Zero Crossing Rate(ZCR): It is defined as a short time ZCR by calculating the weighted average of the number of times the signal[8] changes the sign of the signaling the time slot shown in (6).

$$cr = \sum_{n=1}^{N} \frac{1}{2}[sgn(s_n) - sgn(s_{n-1})] \tag{6}$$

$$where\ sgn(s) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases} \tag{7}$$

iv. Linear Predictive Coding (LPC): this analysis is based on the model of the source filter[20] based on the following transfer function (8)

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{8}$$

Here, ai is the coefficient of the filter. sn is the speech signal and it is assumed not to have any change over analysis frame and is approximately calculated as a linear combination of the previous samples, p as shown in (9)

$$\hat{s}_n = \sum_{i=1}^{p} a_i s_{n-1} \tag{9}$$

In the above equation ai cl be found by minimizing the prediction error of the filter beween sn and šš̂n.

v. Mel Frequency Cepstral Coefficient (MFCC): It resembles human ear hearing characteristics that uses a frequency unit that is nonlinear in nature in order to simulate the auditory system of human. MFCC is calculated by the cosine transform of the logarithm of a short time power spectrum on Mel warped frequency scale.
After framing as done Discrete Fourier Transform is applied in the frames as shown in (10)

$$\mathrm{S[k]} = \sum_{\mathrm{n=0}}^{\mathrm{N-1}} \mathrm{s[n]}.\, \mathrm{e}^{-\frac{\mathrm{j2\pi nk}}{\mathrm{N}}},\ 0 \leq \mathrm{k} \leq \mathrm{N} - 1 \tag{10}$$

Then the Mel filter bank is applied which is basically a collection of triangular filters overlapped where the cut of frequency which is spaced linearly is determined by the central frequency of 2 adjacent filters [23]. These filters are supposed to have a fixed bandwidth on Mel scale. The multiplication is converted into addition by the use of logarithm. The coefficients of the filter bank of the Mel spectrum are calculated as follows (11)

$$F[m] = \log\left(\sum_{k=0}^{N-1} |X[k]|^2 H_m[k]\right), \qquad 0 \leq m \leq M \tag{11}$$

Lastly the Discrete Cosine Transform (DCT) is applied over the log filterbank energies to find the MFCC. (12)

$$c[n] = \sum_{m=0}^{M} F[m] \cos\left(\frac{\pi n(m-1)}{2M}\right),\ 0 \leq n \leq M \tag{12}$$

vi. Normalization of Features: The length of segments are different from each other. Hence to convert them into isometric segments and avoid data redundancy statistical method [14] is used where mean, median, variance, minimum and maximum of the frames are calculated in order to normalize them.

vii. Classifiers Used

a. Support Vector Machines (SVM)

It is a nonlinear classifier which transforms the input feature vector into a high dimension feature space by using a mapping function based on the kernels. Here most of the discrimination is achieved by placing the plane of separation between two class borders in an optimal manner. The plane is extended by Support Vectors thereby reducing the references. Considering P to b linearly separable there will be W ∈ R ^{d} and b ∈ R which satisfies.

$$y_i[(w.x_i) + b] \geq 1, \quad \forall = 1,2,\dots,N \tag{13}$$

Where (w, b) represents a hyper plane.

$$(w.x_i) + b = 0 \tag{14}$$

Given: P= set of points xi ∈RM where i=1…N. The point xi is a component of either of the 2 classes which are labelled as yi ∈ {-1, +1}. The Primary objective is to generate a hyper plane equation which would divide P. The goal requires some primary definitions. Given the set P is separable linearly there is w ∈ Rd and b ∈ R which satisfies

$$y_i[(w.x_i) + b] \geq 1, \quad \forall = 1,2,\dots,N \tag{15}$$

Where (w, b) pair represents a hyper plane.

The objective of finding an optimal hyper plane for separating the classes is converted into a problem using the following:

$$\text{Minimize } W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\text{subject to: } \sum_{i=1}^{N} \alpha_i y = 0, \alpha_i \geq 0, \forall i = 1,2,\dots,N \tag{16}$$

Hence those classifiers which are nonlinear in nature has the potential to become linear by selecting the correct nonlinear Kernels. Kernel functions that are most commonly used are as follows [20].

- Linear kernel

$$K(x_i, x_j) = x_i.x_j \tag{17}$$

- Polynomial kernel

$$K(x_i, x_j) = (x_i.x_j + \beta)^d \tag{18}$$

- Radial Base Function (RBF) kernel

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{19}$$

Generally, for data involving two 2 categories, one single SVM is sufficient for classification. However, since for emotions there are multiple categories of data involved, SVM needs to be generalized for solving multiple class data. There if a problem involving classification has N number of classes, then any 2 different classes can be within the N classes could be classified. However, in such a problem involving N classes, if any 2 classes could be classified then N classes could also be classified by applying combination of rules of Decision Directed Acyclic Graphs (DDAG). It works relatively well when there is a clear margin of separation between the classes.

b. Ensembled Model

The ensembled model which is used in this paper consisted of advanced machine learning algorithms like Random forest classifier and Rotation forest classifier for emotion recognition.

- Random Forest (RF)

  It is a combined algorithm which involved sampling in a random manner and bagging which would generate a specific number of Decision Trees (DT) to develop the forest. It generally selects any random data by replacing training input data in order to generate n number of dissimilar bags to be used for training. Here 100 number of DT are used as the base classifier for evaluating a novel instance. Considering D as the input dataset, then bagging is applied in order to generate n number of dissimilar training datasets (D_i,n) which will be needed for the base classifiers. These dataset bags are then applied with random sampling to create a diversified feature space. The best feature is then selected to split the tree by the DT based on its information gain, Gini Index and gain ratio. Gini Index is used as an evaluation for feature selection that is responsible of classifying the pixels wrongly n comparison to the related classes. Hence, Given, C_i is the targeted class for certain pixel in a random manner and T is equal to the training dataset, then

$$\text{Gini indix} = \sum\sum_{j\neq i}(f(C_i,T)/|T|)(f(C_j,T)/|T|) \quad (20)$$

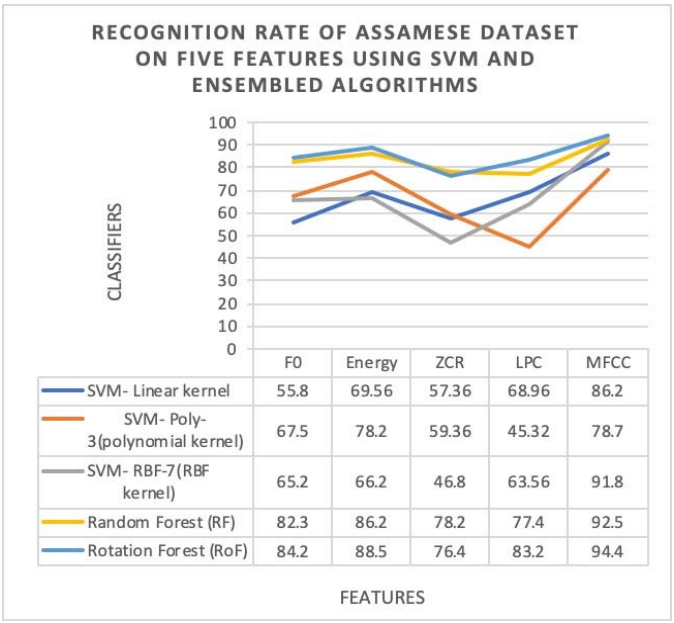
- Rotation Forest (RoF)

  In the RoF algorithm, feature extraction is done randomly and each of the feature set is applied with transformations which are based on decision trees to enhance the accuracy and diversity of the classification involving Ensembled techniques. [26, 31]. The present work with RoF involves application of unsupervised feature extraction method which is Principal Component Analysis (PCA) [28] in order to create a diverse feature space. Attribute bagging is applied to the feature space and the feature space is splitted into subsets which are disjoint in nature having m features. After this new feature set is developed for the base classifier by applying PCA over each of the subset. Here m linear extracted features from each of the subset is combined and these new samples are used for training the Decision tree classifier. Therefore, diversity and precision is achieved in Ensembled algorithm by applying k-axis rotations for the generations of multiple classifier.

3. Results and Discussion

The primary objective of this paper was to evaluate the performance of SVM and Ensembled models under various conditions. The observation was that both RF and RoF achieves better accuracy than SVM. This work involves extensive work of SVM with a variety of kernel functions. The functions used includes linear function, Quadratic function, polynomial function and radial basis functions having k-fold cross validation in performing the experiment. It randomly divides the data into N subsets which are complementary to each other and N-1 subsets are used for training and the remaining on is used for testing. Combination of features were also used to generate the results. Each of the kernel function is utilized in training and testing the SVM on the feature vectors by

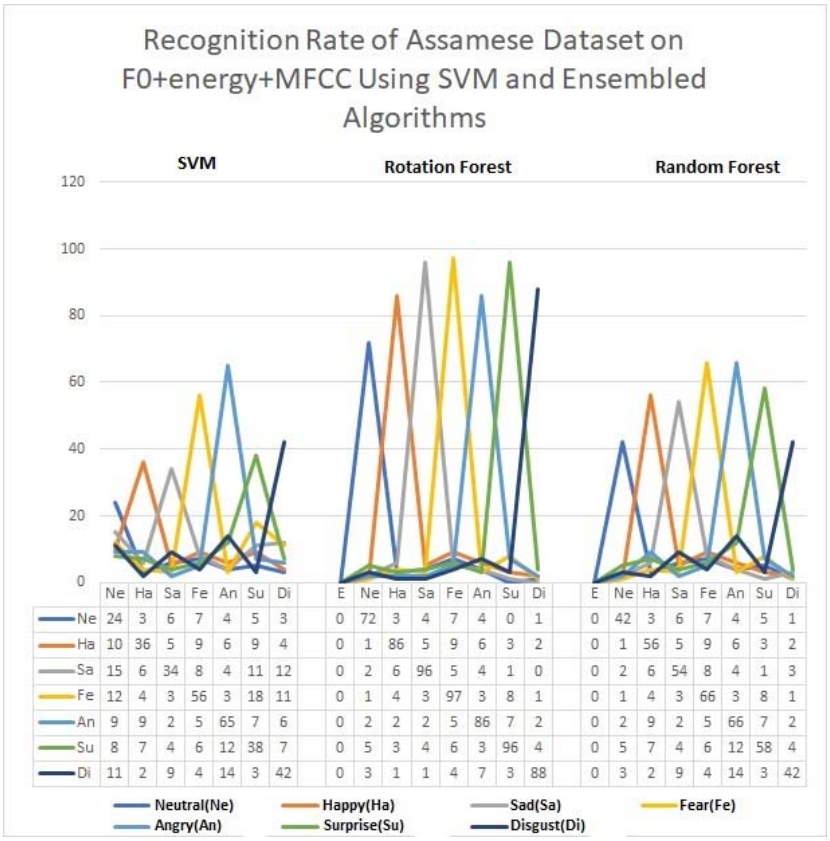
changing the cost values. Linear kernel, poly-3 of polynomial kernel, RBF at sigma 7 are the kernels which gives the best possible results.



	F0	Energy	ZCR	LPC	MFCC
SVM- Linear kernel	55.8	69.56	57.36	68.96	86.2
SVM- Poly-3(polynomial kernel)	67.5	78.2	59.36	45.32	78.7
SVM- RBF-7(RBF kernel)	65.2	66.2	46.8	63.56	91.8
Random Forest (RF)	82.3	86.2	78.2	77.4	92.5
Rotation Forest (RoF)	84.2	88.5	76.4	83.2	94.4

Fig. 3. Accuracy Analysis.

In order to carry out a thorough evaluation of the performance of the experiments were carried out empirically in a variety of feature combinations of the mentioned five features and with all possible kernel functions in SVM. On analysing the results, the combination of features for F0, Energy and MFCC over the linear kernel yields the best possible accuracy of emotion recognition which is of around 94.40%. The confusion matrix is also generated for the same.



	Ne	Ha	Sa	Fe	An	Su	Di	E	Ne	Ha	Sa	Fe	An	Su	Di	E	Ne	Ha	Sa	Fe	An	Su	Di
Ne	24	3	6	7	4	5	3	0	72	3	4	7	4	0	1	0	42	3	6	7	4	5	1
Ha	10	36	5	9	6	9	4	0	1	86	5	9	6	3	2	0	1	56	5	9	6	3	2
Sa	15	6	34	8	4	11	12	0	2	6	96	5	4	1	0	0	2	6	54	8	4	1	3
Fe	12	4	3	56	3	18	11	0	1	4	3	97	3	8	1	0	1	4	3	66	3	8	1
An	9	9	2	5	65	7	6	0	2	2	2	5	86	7	2	0	2	9	2	5	66	7	2
Su	8	7	4	6	12	38	7	0	5	3	4	6	3	96	4	0	5	7	4	6	12	58	4
Di	11	2	9	4	14	3	42	0	3	1	1	4	7	3	88	0	3	2	9	4	14	3	42

Fig. 4. Confusion Matrix of Assamese Dataset on F0+Energy+Mfcc Using SVM & Ensembled Algorithm Recognition Rates.

The experiments were also analysed by parameters for evaluation like Overall Accuracy (OA), Kappa Index Analysis (KIA), Computational Expenses (Trg) and Receiver Operating Characteristic (ROC) for prediction capability. The accuracy of the classifier is evaluated using KIA and OA is responsible for calculating the overall performance of the classifier.

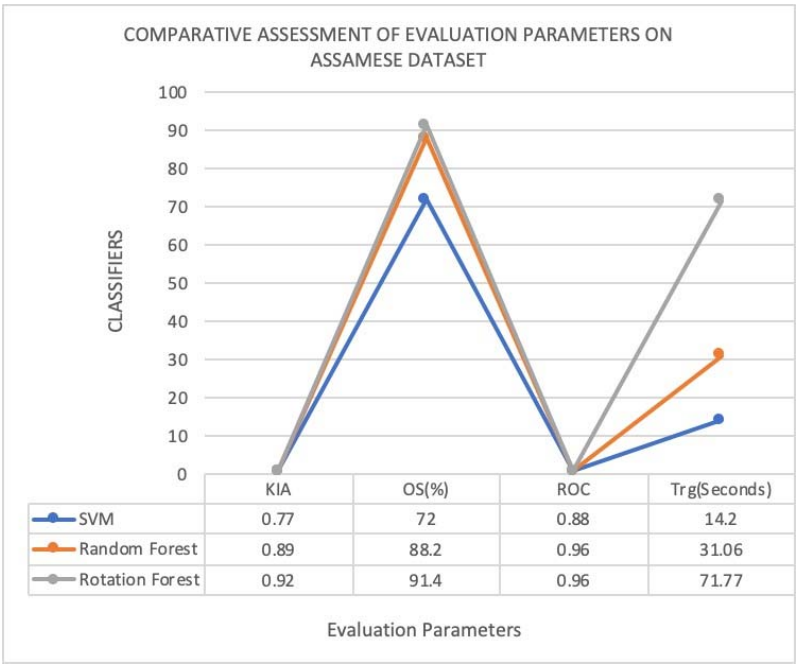


Fig. 5. Comparative Assessment of Evaluation Parameters on Assamese Dataset.

4. Conclusion & Future Scope

This paper deals with the basic feature integration of emotion Recognition using speech. SVM random forest and rotation forest are used to determine the accuracy it is implemented in an entirely new data set which has been generated from the scratch based on the Assamese language and has a combination zero energy etc while for the SVM classifier the linear kernel provides the best possible accuracy for emotion classification and it reports the accuracy to be 91.8% . However for random forest and rotation forest it has been observed that rotation forest gives better result in a combined feature space and insecurity is found to be 92.5% and 94.40% respectively. Using the evolution Matrices it has been observed that the overall performance of rotation forest among all this classifier is the highest however according to the computational expenses it is found to be more expensive discuss the results obtained from this paper can be concluded in several ways which is the emotion recognition has a higher recognition rate when it uses both the prosodic and the spectral features. Secondly the Ensembled algorithms generally has a better performance over the conventional classifier in case of emotion recognition it has also been observed that using combination of a number of features results in in better performance of the classifiers by increasing the accuracy. There is immense scope for improvement of the accuracy of the results using more efficient and modern classifiers like RESNET, Darknet, XceptionNet etc. A real time system can also be developed to give a more practical approach to the research work. ASR integration can also be integrated for the same.

References

[1] I. S. Engberg, and A. V. Hansen, “Documentation of the Danish Emotional Speech Database (DES)”, Internal AAU report, Center for Person Kommunikation, Department of Communication Technology, Institute of Electronic Systems, Aalborg University, Denmark, September 1996.
[2] L. Breiman, “Random forests”, Machine Learning, 2001, vol. 45(1), pp. 5–32, Oct 2001.
[3] Jin, Q., Li, C., Chen, S., Wu, H.: Speech emotion recognition with acoustic and lexical features. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4749-4753 (2015).
[4] F. Yu, E. Chang, Y.Q. Xu, and H.Y. Shum, “Emotion detection from speech to enrich multimedia content”, in Proc. 2nd IEEE Pacific-Rim Conference on Multimedia 2001, pp.550-557, Beijing, China, October 2001.
[5] Lee, J., Tashev, I.: High-level feature representation using recurrent neural network for speech emotion recognition. In: INTERSPEECH (2015).
[6] D. Ververidis, C. Kotropoulos, and I. Pitas, “Automatic emotional speech classification”, in Proc. 2004 IEEE Int. Conf. Acoustics, Speech and Signal Processing, vol. 1, pp. 593-596, Montreal, May 2004.

[7] J. Kreiman and B. R. Gerrat, “Perception of aperiodicity in pathological voice”, Acoustical Society of America, vol.117, pp. 2201-2211, 2005.
[8] L. R. Rabiner and M. R. Sambur, “An algorithm for determining the endpoints of isolated utterances,” Bell System Technical Journal, vol. 54, no. 2, pp. 297-315, February 1975.
[9] Zheng, W. Q., Yu, J. S., Zou, Y. X.: An experimental study of speech emotion recognition based on deep convolutional neural networks. In: International Conference on Affective Computing and Intelligent Interaction, pp. 827-831 (2015).
[10] Chernykh, V., Sterling, G.: Emotion Recognition From Speech With Recurrent Neural Net-works. In: arXiv:1701.08071v1 (2017).
[11] Hermansky, H., “Perceptual linear predictive (PLP) analysis of speech”, The Journal of the Acoustical Society of America, Vol. 87, No. 4, pp. 1738-1752, 1990.
[12] Kim, Y., Lee, H., Provost, E. M.: Deep learning for robust feature generation in audiovisual emotion recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3687-3691 (2013).
[13] Zheng, W. L., Zhu, J., Peng, Y.: EEG-based emotion classification using deep belief net-works. In: IEEE International Conference on Multimedia & Expo, pp. 1-6 (2014).
[14] Han, K., Yu, D., Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine. In: INTERSPEECH (2014).
[15] Fayek, H. M., Lech, M., Cavedon, L.: Towards real-time Speech Emotion Recognition using deep neural networks. In: International Conference on Signal Processing and Communication Systems, pp.1-5, (2015).
[16] Stankovic, I., Karnjanadecha, M., and Delic, V., “Improvement of Thai speech emotion recognition by using face feature analysis”, Proceedings of the Nineteenth IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2011), Chiang Mai, Thailand, December 7-9, pp. 87, 2011.
[17] Dellaert, F., Polzin, T. & Waibel, A., “Recognizing emotion in speech”, Fourth International Conference on Spoken Language Processing, Vol. 3, pp. 1970-1973, Oct. 1996.
[18] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams et al., “Recent advances in deep learning for speech research at Microsoft,” in Proceedings of IEEE ICASSP 2013, 2013.
[19] Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: IEMOCAP: Interactive emotional dyadic motion capture database. In: Journal of Language Resources and Evaluation vol. 42 no. 4, pp. 335-359 (2008).
[20] Nicholson, J., Takahashi, K. & Nakatsu, R., “Emotion recognition in speech using neural networks”, 6th International Conference on Neural Information Processing, Vol. 2, pp. 495–501, 1999.
[21] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B., Zafeiriou, S.: End-to-end multi-modal emotion recognition using deep neural networks. In: IEEE Journal of Selected Topics in Signal Processing (2017).
[22] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. In: EMNLP (2017)
[23] Ranganathan, H., Chakraborty, S., Panchanathan, S.: Multimodal Emotion Recognition Using Deep Learning Architectures. In: Institute of Electrical and Electronics Engineers Inc., United States (2016).
[24] Peipei Shen, Zhou Changjun and Xiong Chen. “Automatic Speech Emotion Recognition Using Support Vector Machine”. Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference, pp.859-862, Aug. 2011.
[25] Richard O. Duda, Peter E. Hart and David G. Stork. “PATTERN CLASSIFICATION”. 2nd ed. New York : Wiley-Interscience, pp.128-138, Oct. 2000.
[26] Milan Sigmund, “Voice Recognition By Computer”, Tectum Verlag publication, pp.20-22.
[27] Douglas Sturim, Pedro Torres-Carrasquillo, Thomas F. Quatieri, Nicolas Malyska and Alan McCree, “Automatic Detection of Depression in Speech Using Gaussian Mixture Modeling with Factor Analysis”, Proceedings of Interspeech, pp. 2981-2984, 2011
[28] Dimitrios Ververidis and Constantine Kotropoulos, “Emotional speech recognition: Resources, features, and methods”, Speech communication, vol. 48, no. 9, pp. 1162-1181, 2016.
[29] Björn Schuller, Gerhard Rigoll and Manfred Lang, “Hidden Markov model-based speech emotion recognition”, Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on. Vol. 1. IEEE, 2003.
[30] JJ. Rodriguez, and LI. Kuncheva, “Rotation forest: A new classifier ensemble method”. IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, pp. 1619–1630, Oct 2006.
[31] Steven R Livingstone & Frank A Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PLOS ONE, Public Library of Science, vol. 13(5), pages 1-35, May 2018.